Istituto di Linguistica Computazionale - Pisa

Grey Literature for Natural Language Processing: a Terminological and Statistical Approach

Laura Cignoni, Gabriella Pardelli, Manuela Sassi
Istituto di Linguistica Computazionale (ILC)
Consiglio Nazionale delle Ricerche (CNR) - Pisa, Italy

Natural Language Processing (NLP) and Computational Linguistics (CL)

Natural Language Processing is *a branch of computer science that studies computer systems for processing natural languages* (Cunningham: 1999)

Computational Linguistics is *a branch of linguistics in which computational techniques and concepts are applied to the elucidation of linguistic and phonetic problems* (Crystal: 1991)

The two expressions are often used indifferently

AIM

Our aim is to contribute to the creation of language resources of grey literature terms of the last decades. This can help prevent the disappearance of documents containing words that have undergone rapid changes and that represent the main anchors for information retrieval

Statistical representation

The most significant old and new terms relative to grey literature in the field of natural language processing and other interrelated disciplines have been associated, highlighting the terminological changes that have taken place in the course of time

ROLE OF TERMINOLOGY

As the queries are often incorrect, inappropriate, or simply far too general, it is necessary to integrate pre-existing or obsolete words and expressions used by specialists in the different domains to create a synonym relationship between the terms contained in the different NLP documents. In this way a term, even if dated and no longer in use, can become the key to enter the world of knowledge

GREY LITERATURE CORPUS

Our grey literature corpus is composed of ca 13,000 records corresponding to the titles of papers presented at International Conferences in the field of natural language processing (1950 to June 2008)

SOURCES

The main sources for our Corpus include:

- ACL Anthology
- LREC Conferences
- Weaver Memorial
- Alpac Report
- Conferences on Automatic Translation

Methodology

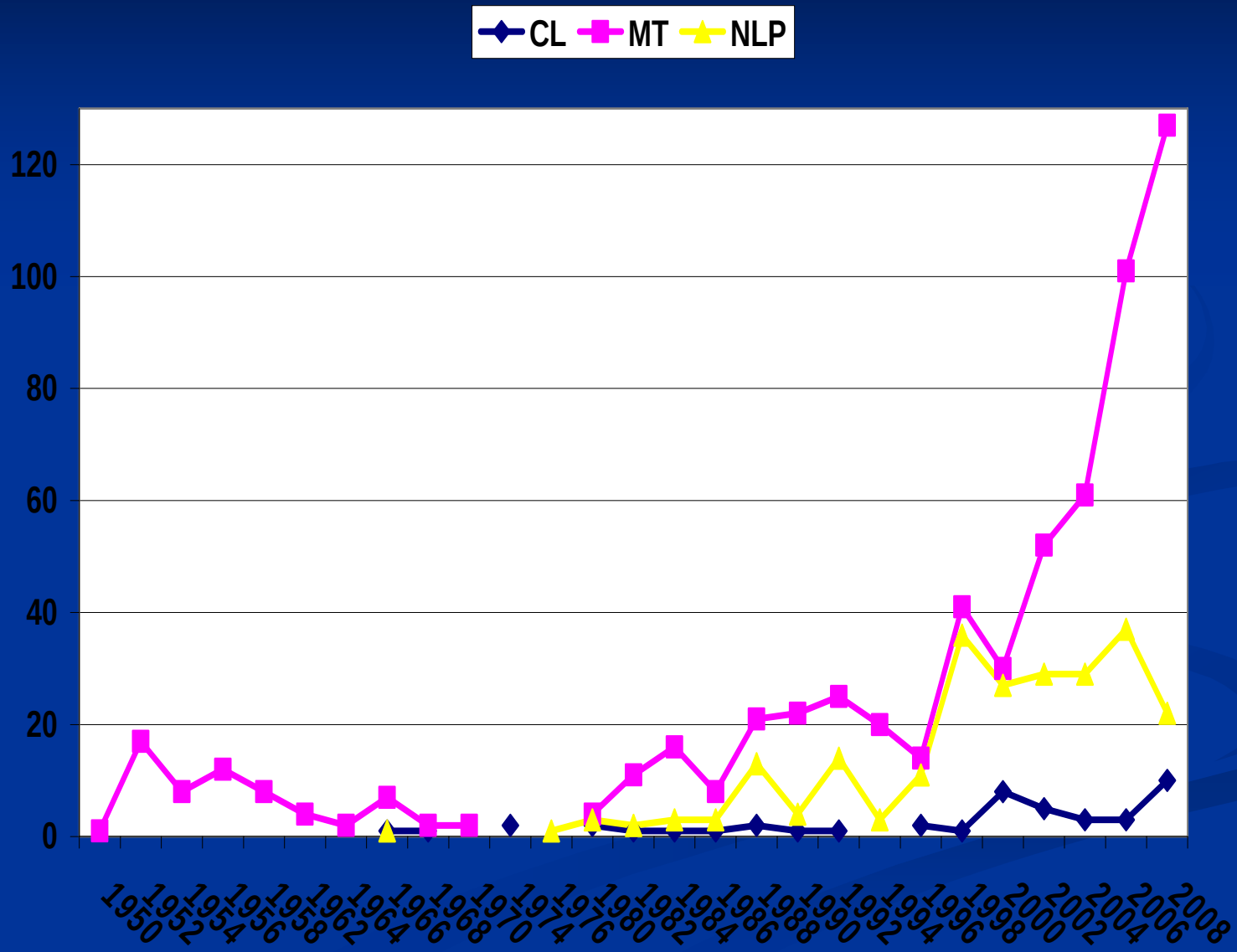
The methodology used is the following:

- Search and saving of the most common single terms which are the object of this study
- Extraction of the contexts with year and abbreviation of the conference
- Generation of tables according to the chronological use of these terms
- Creation of charts

Words extracted from GL Corpus

automated, automatic, automatically, automatically-extracted, automating, automation, automatique, automatisation, automatischen, automatisée, automatism, automatized, computability, computation, computationally, computational, computationally, computational-semantic, computations, compute, computed, computer, computer-aided, computer-assisted, computer-based, computerization, computerized, computer-mediated, computers, computes, computing, mechanical, mechanized, machina, machine, machine-aided, machine-guided, machine-induced, machine-learning, machine-mediated, machine-readable, machines, machine-tractable, machine-translation, electronic translation

Trend of Computational Linguistics (CL), Machine Translation (MT) and Natural Language Processing (NLP)



MOST FREQUENT CO-OCCURRENCES

DS	*Dialogue System(s)*
IE	*Information Extraction*
IR	*Information Retrieval*
LG	*Language Generation*
LR	*Language Resource(s)*
ML	*Machine Learning*
NE	*Named Entity*
PC	*Parallel Corpora*
QA	*Question Answering*
SD	*Spoken Dialogue(s)*
SR	*Recognition Speech*
WSD	*Word Sense Disambiguation*

Use of co-occurrences